Price and Popularity: A Data Science Approach to Steam Games

Capstone Project for CSCI1951A: Data Science

Yuchen (Juliana) Han

Steam is the most popular video game digital distribution service. Sparked by a discussion on notable disparities in game prices across diverse countries, my team of four embarked on an exploration of the pricing strategies of Steam games. We initially hypothesized that a country's purchasing power could have an impact on game prices. However, as we delved deeper into our investigation, we discovered that several other factors also impact game prices and popularity including genres, ratings, publishers, etc.

Our analysis combined three datasets for our analysis: firstly, a Kaggle dataset featuring top 3000 games with attributes such as price, popularity, ratings, genres, publishers; secondly, price data for these games across six selected countries, obtained through Steam's official API; and thirdly, the Consumer Price Index (CPI) of the six countries, sourced from International Monetary Fund. We performed data cleaning and processing for streamlined data analysis.

We performed three statistical tests for hypotheses on the factors influencing game prices and the impact of price on other parameters. We also implemented three machine learning models to examine the interrelationships across multiple attributes more comprehensively.

Statistical Tests Results

Hypothesis	Country CPI and game prices across countries ar4e positively correlated.	The average prices of games across the top 10 genres are different.	Free-to-play games have more daily active users than paid games.
Test	Pearson ρ	ANOVA	Mann-Whitney U
p-value	p = 0.031	p = 3.335e-84	p = 0.003
Decision	Reject Null	Reject Null	Reject Null

Machine Learning Model Results

As we hypothesize that game popularity is affected by various features, we first built a polynomial Ridge regression to learn such correlation. However, the complexity of the dataset,

as indicated by a negative R-squared value, suggested the need for a high-degree polynomial regression model, which was beyond our resource capacity. To improve model performance, we trained a 3-layer neural network, achieving a test accuracy of 82.16%. To further generalize our findings, we developed a hierarchical clustering algorithm to explore latent patterns in game attributes, and presented the results visually using Principal Component Analysis (PCA).

Interactive Component

Our interface offers a dynamic user experience with selectable 1D, 2D, or 3D visualizations across five cluster number options, allowing users to interact with cluster and game data. Additionally, a game title search feature is included, providing details of the corresponding cluster and other relevant game information.